DOCUMENT RESUME

ED 159 190 TM 007 382

AUTHOR Clark, Mary Jo; Blackburn, Robert T.
TITLE Assessing Faculty Performance: A Test of Method.
PUB DATE 31 Mar 78
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS College Faculty; *College Teachers; *Effective Teaching; Higher Education; Peer Evaluation; *Rating Scales; Self Evaluation; Student Evaluation of Teacher Performance; Teacher Administrator Relationship; *Teacher Evaluation; *Test Reliability; *Test Validity

ABSTRACT

A methodology for evaluating faculty work performance was discussed, using data obtained from a typical liberal arts college faculty. Separate evaluations of teaching effectiveness and of overall contributions to the college for 45 full-time faculty (85% response rate) were collected from administrators, faculty colleagues, students, and from the professors themselves. The reliability and validity of quality ratings followed Cartter (1966). Independent questionnaire data on demographic variables, professional attitudes and opinions, personality measures, job satisfaction, and perceived stress in the work situation gave correlation coefficients whose sign was predicted from role theory. These measures added to construct validity. The procedure satisfied statistical tests for reliability and validity. Practical suggestions for use of this method are included. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159190

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mary Jo Clark

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ASSESSING FACULTY PERFORMANCE: A TEST OF METHOD

Mary Jo Clark
Research Psychologist
Educational Testing Service
Princeton

Robert T. Blackburn
Professor of Higher Education
Center for the Study of Higher Education
University of Michigan
Ann Arbor

Draft for presentation at the annual meeting of the American Educational Research Association on March 3[illegible], 1978, in Toronto, Ontario, Canada.

TM007 382

ASSESSING FACULTY PERFORMANCE: A TEST OF METHOD

ABSTRACT

The study introduces a methodology for evaluating faculty work performance. Data are obtained on a typical liberal arts college faculty. Separate evaluations of teaching effectiveness and on overall contributions to the college for 45 full-time faculty (85% response rate) were collected from administrators, faculty colleagues, students, and from the professors themselves.

The reliability and validity of quality ratings follow Cartter (1966). Independent questionnaire data on demographic variables, professional attitudes and opinions, personality measures, job satisfaction, and perceived stress in the work situation gave correlation coefficients whose sign was predicted from role theory. These measures added to construct validity.

The procedure satisfied statistical tests for reliability and validity. Practical consequences are inferred.

INTRODUCTION

Reputational ratings of the quality of American colleges and of university departments began with R. M. Hughes in 1925 and have continued at irregular intervals in the intervening years. With every published report, questions have been raised regarding the reliability and validity of the rankings that inevitably result. Methodological concerns have continued to be raised about the most recent assessments, such as the one conducted by Harnett, Clark, and Baird (1978), despite the fact that Cartter, in his Assessment of Quality in Graduate Education (1966), presented persuasive evidence that expert ratings are quite consistent with other measures of program quality.

The essence of Cartter's (and others') methodology is the averaging of expert judgments or ratings made by professional peers, and the use of a single global assessment rather than ratings based on specific program characteristics. Cartter gave considerable attention to the question of the reliability (and by inference, the construct validity) of subjective evaluations concerning the concept, "quality." He noted that "in an operational sense, quality is someone's subjective assessment, for there is no way of objectively measuring what is in essence an attribute of value (1966:4)." Comparing the subjective evaluations of quality with data about other program characteristics that are frequently used in academic evaluation, Cartter then stated: "We conclude from these various comparisons that a survey based on the opinions of well-informed scholars within the academic community is as reliable a guide as one can devise in attempting to measure the elusive attribute of quality (1966:9)."

While the debates on assessing departments continue, the past few years have witnessed the introduction of a similar problem at the level of the assessment of faculty performance. Attention to such assessment accelerated with student demands for better information about the teaching performance of faculty members. With increased external demands for accountability and internal pressures for fairness because of limits set on promotion and tenure, the need to have quantitative assessments by peers and administrators as well as by students has been recognized. The problem of assessing faculty performance is both delicate and complex, for obvious reasons. Moreover, the first studies using colleague ratings had very low reliabilities for specific items on the evaluation instrument. (See, e.g., Centra (1975).) Hence, much needs to be known in this area in order that faculty fears can be allayed and that sound decisions affecting the welfare of higher education be made.

This study provides data on the reliability and validity of using global ratings to assess faculty performance. How reliable are ratings made by students, faculty colleagues, administrators, and self, and how well do such ratings agree with one another?

THE SETTING

All data reported here were collected from students and faculty members at a small liberal arts college in the Midwest. "Midwest" College's 1500 undergraduate students encompassed a full range of academic interests and qualifications. A moderately well-trained faculty were spread over the typical departments. The principal faculty roles were teaching and contributing to the institution (i.e., committee work, advising, program development) with little demand for or expectation of

published scholarship. In these and other ways, Midwest College is very near the middle in comparison with about 800 other private and church-related liberal arts colleges. With respect to activities in the faculty role, Midwest is not unlike many state colleges or universities except, of course, Midwest is much smaller.

PROCEDURES

For a number of reasons, Midwest faculty willingly participated in a series of data collection procedures that were part of a wider-ranging college self-study. Forty-five of the 53 full-time faculty (85%) provided full information. Among other things, faculty members rated their colleagues as well as themselves on teaching effectiveness and on overall contribution to the college. Midwest's administrators (president, academic dean, and department heads) also rated faculty members on both dimensions of performance. Course evaluations obtained from students in the previous semester provided a third measure of teaching effectiveness. A sample of faculty participated in a second set of colleague ratings to assist in the estimation of measurement reliability.

Specifically, each faculty member rated every other teacher in her or his curricular division and herself or himself on a five-point scale of "teaching effectiveness." Administrators also rated faculty members on this dimension. The rater was told to "consider those qualities which are important in the evaluation of the skills and practices and products of a classroom teacher regardless of rank or experience or training of the person being rated." In a similar way, each faculty member judged self and colleagues on a five-point scale concerning "overall contribution" to Midwest College. Again, administrators also rated faculty

members. The rater was told to "take into account the person's total contribution, whether his own work or his stimulation of others, whether scholarly or administrative or in human relations; the person's overall usefulness in helping the college carry out its responsibilities."

Student evaluations of teaching effectiveness were obtained from a standard 14 item five-point scale questionnaire to evaluate all courses each semester. Responses to the question "how would you rate your instructor in teaching effectiveness?" were averaged across all courses taught by a faculty member during one semester for an index of teaching performance as judged by students.

These performance measures reflect the effect of behavior on other persons or situations rather than specific behaviors themselves. That is, ratings on teaching effectiveness and overall contribution to the college are expert judgments by colleagues and students based upon unspecified but surely varied observations and sources of information. They are value judgments of the "effectiveness" of outcomes, not judgments about the specific behaviors of the individual faculty members. The use of expert judgments as performance measures in this study assumed that the results of faculty role performance are apparent to colleagues, and that this is particularly true in a small college setting where an individual's work becomes known to most others. Furthermore, it was assumed that the opinions of colleagues and students provided salient feedback for the faculty member as he or she attempted to perform effectively as a college professor. Therefore, the consistency of ratings across raters with different sources of information and perspectives would be a measure of the validity of the responses as performance criteria.

The argument builds as follows: Professors have professional norms as a common general base for judgments about the effective performance of their peers. Faculty members in the same or related academic disciplines have some common understanding of desirable teaching outcomes in their areas. Students are observers and also the objects of teaching performance. Administrators make decisions for the overall welfare of the institution based on judgments about the performance of its members. Faculty members and administrators are professionally trained in their disciplines to make judgments based on evidence and objectivity. Finally, at least in a comparatively intimate setting, faculty members, administrators, and students are in close proximity and have considerable opportunity to observe and know each other. For all of these reasons, ratings on "effectiveness" or "contribution" made by these judges meet the requirement of logical relevance with regard to the concept of "effective" faculty role performance.

Nevertheless, problems with rating scales do exist. Errors of leniency, errors of central tendency, biases of selective perception, halo effect, different philosophical bases or beliefs about the desirability of certain behaviors, and the lack of opportunity for first-hand observation have been investigated and reported in the research literature (Guilford, 1954; Remmers, 1963; Biddle and Ellena, 1964; Torrence, 1965). All of these possibilities for systematic error threaten the confidence that ratings really represent what they are said to represent. Such errors cannot be avoided entirely but they can be reduced by selecting raters with access to relevant information on which to base judgments, selecting raters with a common frame of reference in which to make judgments, and structuring the rating task with a minimum of ambiguity so that raters

are most likely to interpret it similarly. In the evaluations collected at Midwest College, all of these methods were used to minimize variation that might be due to the biases of raters rather than the characteristics of the persons being rated.

Concerning judgments about the teaching effectiveness of faculty members, the lack of opportunity for first-hand observation in the classroom has been one of the most persistent challenges to the validity of peer ratings. But professors frequently exchange information about their courses and have a number of opportunities to observe one another's work. In addition, when judgments are made by general impression (rather than by analysis of specific elements), some of the "halo" effects reflect actual variations in the information available to the different judges rather than errors to the judging operation itself. Thus, although judgments by general impression may be less consistent than more analytical judgments, they are not necessarily less valid.

There is no question that there is some ambiguity in the task of rating "teaching effectiveness" and "overall contribution to the college," and therefore responses will reflect some of the styles and views of the individual raters. However, these differences among professionals in the same field are always part of judgments being made about each other. It would be unrealistic to try to remove all such differences in perspective and judgment; variations in assessments of the professional role performance of faculty members need to be taken into account when making college or university decisions, but do not necessarily invalidate the judgment procedure.

Reliability of the Ratings

A random sample of nine faculty members completed a second set of peer ratings on teaching effectiveness and overall contribution to the college a few weeks after the first collection of colleague ratings, in order to obtain an estimate of the stability of the ratings. The test-retest correlations on rated teaching effectiveness ranged from .75 to .88 with an average estimated reliability coefficient of .81. Correlations between ratings on overall contribution to the college at time one and time two were somewhat more variable, ranging from .23 to 1.00 with an average estimated reliability of .61. Though based on a small sample, these test-retest reliabilities demonstrate a considerable amount of stability in the evaluations of peers made by the faculty members at Midwest College.

Following the procedures outlined by Ebel (1951) and Guilford (1954), an intraclass correlation was used to estimate the inter-rater agreement in mean judgments made by faculty raters about their colleagues. This procedure tends to give a fairer estimate of internal consistency than does the general Spearman-Brown reliability formula when the rating indices are based on different numbers of individual ratings, as in this study (Ebel, 1951). Using a multifactor one-way analysis of variance to obtain estimates of person and error variance, the estimated inter-rater reliability of mean rated teaching effectiveness is .72 and the inter-rater reliability of mean rated overall contribution to the college is .86. The reliability estimate for rated teaching effectiveness is based on a median of four ratings per person and the estimate for rated overall contribution to the college is based on an average of 30 ratings per person. Guilford (1954) states that reliabilities for ratings often run around .75.

These data are consistent with his findings.

Separate evaluations of inter-rater reliability were not carried out for ratings made by administrators or students, because of limitations in the available data. Probably the intraclass correlation of mean ratings by administrators would be somewhat lower since the average number of ratings per person is lower (two ratings on teaching effectiveness and three on overall contribution). The inter-rater reliability for mean student rated teaching effectiveness should be higher because of the large number of ratings combined to form this index. High reliabilities for student ratings have been reported since the early work of Shock, Kelly, and Remmers (1927) to recent and current reviews of the literature by Costin, Greenough, and Menges (1971) and by Feldman (1977). High correlations between student used forms have also been found (Sherman and Blackburn, 1975:127). Obviously the self-ratings can only be single judgments, and therefore cannot be evaluated for reliability with the usual statistical procedures.

Validity of the Ratings

In addition to ratings on the two dimensions of role performance, each faculty member in the study completed a questionnaire on background information, selected personality dimensions, job satisfactions, professional attitudes and opinions, and perceived stress in the work situation. The questionnaire was constructed to measure a number of independent and intervening variables in the larger research design, based on role theory and previous studies of organizational stress (Kahn, et al., 1964). To evaluate the performance criteria, items from the questionnaire were predicted to correlate positively or negatively with mean rated overall contribution to the college. Predictions were based on relationships suggested by

theory, experience, and the research literature. Agreement between performance ratings and these characteristics provide a test for construct validity (Cronbach and Meehl, 1955). Although high correlations were not anticipated, it was hypothesized that consistent patterns of correlation in the expected direction would support the claim of rating validity.

Table 1 presents the proportion of agreement for each of the four conditions by each of the four rater groups. More than two-thirds of the correlations are in the predicted direction for all but three cells.

[Insert Table 1 about here.]

Four cells reach 88% agreement. Only mean administrator ratings on teaching effectiveness fail to correlate as predicted with a majority of the items.

A few examples illustrate the patterns and demonstrate the complexity of the relationships. For instance, expressed satisfaction with one's work is significantly and positively related to one's self-rating on teaching effectiveness but shows a low negative correlation with mean teacher effectiveness as rated by each of the other reference groups. Having a higher academic rank is significantly related to mean rated teaching effectiveness by colleagues but not by administrators, students, or self. Years of experience carries most weight in administrator ratings. Publications are a factor for peers and students but not for administrators or self. Self-esteem is related to perception of teaching effectiveness by students and self but not by peers or administrators. All rater groups demonstrate a relationship between seeing that one's work is carefully planned and organized and being an effective teacher; the

correlation is highest with self-ratings. On the other hand, only self-ratings suffer from feelings of pressure because of not measuring up to the demands of the job, and only self-ratings show a positive relationship with teaching load. Committee service is related to mean rated overall contribution to the college for all groups of raters but satisfaction with one's type of work shows a significant relationship only with self-rated overall contribution. These variations in the strength of relationship depending upon the rater group make the predicted pattern of positive and negative correlations across rater groups even more impressive.

Two performance measures and four groups of raters also permit use of the method for assessing convergent and discriminant validity suggested by Campbell and Fiske (1959). The performance dimensions of teaching effectiveness and overall contribution to the college serve as multiple traits being measured. The different groups of raters become multiple methods of measurement. The intercorrelation matrix presented in Table 2 then provides a multitrait-multimethod test of measurement validity.

[Insert Table 2 about here.]

Convergent validation compares agreement in trait measurement across independent groups of raters. Discriminate validation asks whether there is a real difference between the performance dimensions being rated. To the extent that agreement is attained among groups on the way they rate faculty members on each performance dimension, similar properties of the dimension rather than particular interpretations or biases of each rater group are demonstrated. Different ratings on the two performance dimensions by the same group of raters demonstrates discrimination between traits.

The validity coefficients of particular interest in this analysis are underlined in Table 2. Campbell and Fiske (1959) suggest three ways to check for validity. First, the correlations in the validity diagonals should be significantly different from zero to demonstrate agreement on a trait across groups of raters. In Table 2 it is seen that rated teaching effectiveness by faculty colleagues correlates significantly with similar ratings by administrators and students. The correlation between administrator and student ratings on teaching effectiveness, though lower, is also statistically significant. But self-ratings on teaching effectiveness have near-zero correlations with ratings by the other three groups of raters. Among self-ratings, only colleague ratings on overall contribution to the college demonstrate a statistically significant relationship, and then only a moderate one (0.45).

Second, correlations in the validity diagonals should be higher than other values in the same column or row to demonstrate more agreement across groups of raters on the same dimension than on different dimensions. Again, only self-ratings fail to meet this test. The correlation between self-rated teaching effectiveness and colleague-rated teaching effectiveness is lower than correlations between self-rated teaching effectiveness and colleague-rated overall contribution. There is little commonality between self-perceptions of teaching effectiveness and judgments on this same dimension by colleagues, students, and administrators. There is, however, fair agreement among the three independent groups of raters. There is agreement between self and colleague ratings on overall contribution to the college. However, the correlation is very low between self and administrator ratings on this dimension.

Third, to demonstrate discrimination between ratings on teaching

effectiveness and on overall contribution to the college, the validity coefficients should be higher than the correlation of the two traits within each rater group. From Table 2 it can be seen that mean ratings made by colleagues and administrators on each dimension correlate more highly (.625 and .544) than correlations between the two dimensions within each group (.450 and .429). Although the two dimensions are related, these rater groups appear to be making consistent distinctions between them. But this is not the case for self-ratings, where internal agreement (.723) is much higher than agreement with the other raters on either performance dimension. Professors appear to view their own teaching effectiveness and overall contribution to the college in nearly interchangeable ways, even though they seem able to make distinctions between the two performance criteria when rating their colleagues.

Another way to view the extent to which the judgments of different sets of raters are measures of the same underlying concept--in this case, teaching effectiveness or overall contribution to the college--is to compute the common factor reliability of each set of ratings in relation to the concept (Werts, Linn, and Jöreskog, 1974). This procedure starts by assuming that there is a single common or true factor underlying the ratings made by different rater groups, and uses the correlation coefficients between any three sets of ratings (see Table 2) to estimate how well each of them measures the common factor. The assumption of a common factor is supported if reasonably high reliabilities are obtained for all three sets of ratings. It is not supported if a very low estimate of reliability is obtained for one or more of the sets of ratings, or if the results are logically impossible (e.g., reliabilities greater than one).

Applying the procedure to the student, faculty, and administrator ratings of teaching effectiveness obtained in this study, the estimated reliability of the ratings by each group are, respectively, .68, .72, and .69. These moderately sized reliabilities are consistent with the assumption of a single underlying factor, "teaching effectiveness." Since all three sets of ratings appear to be measuring substantially the same concept, a composite of the three ratings would be even more reliable than are the ratings made by any one group.

The self-ratings of teaching effectiveness do not appear to be a good measure of what is being measured by ratings by students and faculty since the estimate of common factor reliability for self-ratings is only .30. There is even less agreement between the self-ratings on overall contribution to the college and the ratings made by faculty members and administrators on this performance dimension. The hypothesis that faculty, administrator, and self-ratings on overall contribution to the college are measures of the same concept must be rejected since the computations result in the logical impossibility of a reliability coefficient greater than 1.0. This results largely from the low level of association between self-ratings and the other two measures.

These analyses suggest that faculty members evaluate their own professional performance quite differently from the way it is evaluated by their colleagues on the faculty, their students, and especially by administrators. The implications of this outcome are explored in Blackburn and Clark, 1975. However, the analyses provide strong support for the hypothesis that global ratings of teaching effectiveness made by faculty colleagues, administrators, and students represent different views of one common underlying factor.

CONCLUSIONS

Conclusions from the data are:

1. Global ratings by colleagues and students on relatively undefined dimensions such as "teaching effectiveness" and "overall contribution to the college" are reasonably reliable indicators of these dimensions. That the global rather than trait specific measures achieve acceptable reliability levels is partly explained by Nisbett and Wilson (1977:244-5). They demonstrated that a student judgment of a professor was first made on a general impression and later defended on rating particularistic attributes, a reverse causal sequence of events the subjects erroneously deny they used. (Multidimensional evaluations can continue to provide important data for identifying strengths and weaknesses--for specific teaching arts or for program review, for example.)
2. Administrator ratings probably also meet these tests, but the evidence is less conclusive. Self-ratings on these dimensions appear to mean something different than the global ratings made by others.
3. Apparently, performance judgments as made by colleagues, students, administrators, and self, reflect different sources of information or points of view. The results suggest that more effort should be made to obtain general performance appraisals from faculty colleagues and probably also from administrators, in addition to course and teacher evaluations from students, when making judgments about faculty job performance. Multiple measures benefit from different perspectives concerning the

aspect of work that is being evaluated; combining the measures would lead to more reliable assessment than using any one measure alone.

4. Self-ratings on general factors such as "teaching effectiveness" or "overall contribution to the college" do not appear to mean the same thing as ratings by others on these factors; therefore, such self-ratings probably should not be combined with the ratings of others when conducting performance appraisals. This does not, however, negate the value and importance of clear communication to faculty members concerning their work performance. In addition to feedback from faculty colleagues and administrators as well as from students, the individual should be involved in specific discussions of ways to strengthen job performance.

The results of this study need corroboration in other settings, particularly larger universities where decentralization may create isolated clusters of faculty members and possibly inadequate peer knowledge of performance. Also, the use of global ratings to assess the performance of faculty members with research as a principal activity needs to be considered. The research role also appears to be appropriate for assessment by this methodology.

REFERENCES

Biddle, B. J., & Ellena, W. D. (eds.). Contemporary research on teacher effectiveness, New York: Holt, Rinehart & Winston, 1964.

**Blackburn & Clark

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-104

Cartter, A. M. An assessment of quality in graduate education. Washington, D.C.: American Council on Education, 1966.

Centra, J. A. Colleagues as raters of classroom instruction. Journal of Higher Education, 1975, 46, 327-337.

Costin, F., Grennough, W. T., & Menges, R. T.. Student ratings of college teaching: reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-535.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-320.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Feldman, K. A. Consistency and variability among college students in rating their teachers and courses: a review and analysis. Research in Higher Education, 1977, 6, 223-274.

Guilford, J. P. Psychometric methods (2nd Ed.). New York: McGraw-Hill, 1954.

Hartnett, R. T., Clark, M. J., & Baird, L. L. Reputational ratings of doctoral program. Science, 24 March, 1978, 199, 1310-1314.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., & Snoek, J. D. Organizational stress: Studies in role conflict and ambiguity. New York: Wiley, 1964.

**Blackburn, R. T., & Clark, M. J. An assessment of faculty performance. Sociology of Education, 1975, 48, 242-256.

Nisbett, R. E., & Wilson, T. D. Telling more than we can know: verbal reports on mental processes. Psychological Review, 1977, 84, 231-259.

Remmers, H. H. Rating methods in research on teaching. In Gage, N. L. (ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Sherman, B. R., & Blackburn, R. T. Personal characteristics and teaching effectiveness of college faculty. Journal of Educational Psychology, 1975, 67, 124-131.

Shock, N. W., Kelly, E. G., & Remmers, H. H. An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue rating scale. Journal of Educational Psychology, 1927, 18, 187-195.

Torrence, E. Constructive behavior: Stress, personality, and mental health. Belmont, Calif.: Wadsworth, 1965.

Voeks, V. W. Publications and teaching effectiveness. Journal of Higher Education, 1962, 33, 212-218.

Werts, C. E., Linn, R. L., & Jöreskog, K. G. Intraclass reliability estimates: Testing structural assumptions. Educational and Psychological Measurement, 1974, 34, 25-33.

TABLE 1

PROPORTION OF CONFIRMED POSITIVE OR NEGATIVE CORRELATIONS BETWEEN ITEM RESPONSES AND MEAN RATED ROLE PERFORMANCE

Item prediction	Percent of agreement with mean rating by			
	Peers	Administrators	Students	Self
Positive correlation with rated teaching effectiveness (18 items)	83	44	78	83
Negative correlation with rated teaching effectiveness (15 items)	67	67	67	67
Positive correlation with rated overall contribution (17 items)	88	59	--	88
Negative correlation with rated overall contribution (8 items)	63	88	--	88

TABLE 2

RATED TEACHING EFFECTIVENESS AND RATED OVERALL CONTRIBUTION TO THE COLLEGE AS EVALUATED BY PROFESSORS, ADMINISTRATORS, SELF, AND STUDENTS IN CLASSES

	Professors		Administrators		Self		Students
	Teaching (N=45)	Contribution (N=45)	Teaching[a] (N=29)	Contribution[b] (N=45)	Teaching (N=40)	Contribution (N=41)	Teaching[c] (N=45)
Professors							
Teaching	(.72)						
Overall Contribution	.488*	(.86)					
Administrators							
Teaching[a]	.625*	.243	( )				
Overall Contribution[b]	.165	.544*	.429*	( )			
Self							
Teaching	.278	.335*	.098	.126	( )		
Overall Contribution	.333*	.450*	.065	.146	.723*	( )	
Students							
Teaching[c]	.620*	.240	.466*	-.036	.187	-.070	( )

* Correlations are significantly different from zero at or above the 95% level of confidence.

a. The administrator rating on teaching effectiveness is the mean of ratings by the appropriate division chairman and by the academic dean.

b. The administrator rating on overall contribution to the college is the mean of ratings by the president, academic dean, and assistant dean.

c. The student rating is a mean of course evaluation responses to the question: "How would you rate your instructor in teaching effectiveness?"